9-2015

Informational content of factor structures in simultaneous discrete response models

Shakeeb KHAN

Arnaud MAUREL

Yichong ZHANG
Singapore Management University, yczhang@smu.edu.sg

Informational Content of Factor Structures in Simultaneous Discrete Response Models

S. Khan
Duke University

A. Maurel
Duke University and NBER

Y. Zhang
Duke University

September 19, 2015
Preliminary and Incomplete Version

Abstract

We study the informational content of factor structures in discrete triangular systems. Factor structures have been employed in a variety of settings in cross sectional and panel data models, and in this paper we attempt to formally quantify their informational content in a bivariate system often employed in the treatment effects literature. Our main findings are that under the factor structures often imposed in the literature, point identification of parameters of interest, such as both the treatment effect and the factor load, is attainable under weaker assumptions than usually required in these systems. For example, we show is that an exclusion restriction, requiring an explanatory variable in the outcome equation not present in the treatment equation is no longer necessary for identification. Furthermore, we show support conditions of included instruments in the outcome equation can be substantially weakened, resulting in settings where the identification results become *regular*. Under such settings we propose a estimators for the treatment effect parameter, the factor load, and the average structural function that are root-n consistent and asymptotically normal. The estimators' finite sample properties are demonstrated through a simulation study and in an empirical application, where we implement our method to the estimation of the civic returns to college, revisiting the work by Dee (2004).

Keywords: Factor Structures, Discrete Choice, Treatment Effects.

1 Introduction

Factor models and factor structures continue to see widespread and increasing using in various areas of econometrics. This type of structure has been employed in a variety of settings in cross sectional, panel and time series models, and have proven to be a flexible way to model the behavior of and relationship between unobserved components of complicated models. Furthermore, they have shown to facilitate the identification of structural parameters without the need for stringent parametric specifications, allow for a data driven way reduce the dimensionality of a given semiparametric or nonparametric model, and reduce the reliance on exclusion restrictions in nonlinear simultaneous equation models.

The general idea behind factor models is to assume that the dependence across the unobservables is generated by a low-dimensional set of mutually independent random variables (or factors). The applied and theoretical research in econometrics employing factor structures is extensive. These models are typically used in the treatment effect literature as a way to identity the joint distribution of potential outcomes from the marginals, and then recover the distribution of treatment effects from this joint distribution.

Factor models have become increasingly popular in recent years. Recent papers using these models in the context of treatment effect estimation includes, among many others, Carneiro, Hansen, and Heckman (2003), Aakvik, Heckman, and Vytlacil (2005), Cunha, Heckman, and Navarro (2005), Cunha and Heckman (2007), Heckman and Navarro (2007), Cooley, Navarro, and Takahashi (2015). See also Abbring and Heckman (2007) for a more extensive list of references.

Factor models have been used in a number of other contexts in economics. Notably, factor models have also been used in the context of earnings dynamics (see, e.g., Abowd and Card (1989), Horowitz and Markatou (1996), Bonhomme and Robin (2010)) as well as cognitive and non-cognitive skill production technology (Cunha, Heckman, and Schennach (2010)). All of these papers, with the notable exception of Cunha, Heckman, and Schennach (2010), rely on linear factor models where the unobservables are assumed to be given by the sum of a linear combination of mutually independent factors and an idiosyncratic shock.

Factor structures are also used in financial econometrics. In these settings it is shown that factor structures can allow for new models for the dependence structure, or copula, of economic variables based on a latent factor structure. This can be particularly attractive for

relatively high dimensional applications, involving dozens or more variables- see for example, Oh and Patton (2013),Oh and Patton (2015), Hull and White (2004). In panel data models factor models have also been useful too allow for more general forms of nonstationarity and dynamics- see, e.g. Bai and Ng (2002),Khan, Ponomareva, and Tamer (2015). In time series models with factor structures, see, e.g. Stock and Watson (2011).

In this paper we explore the *informational content* of factor structures often employed, in a particular class of models. This class can best be described as a system of simultaneous discrete equations. Focusing on this class of models can be well motivated from both an applied/empirical and theoretical perspective. From the former, many treatment effect models and parameters of interest fit into this framework as treatment is often a binary and endogenous variable in the system, whose effect on outcomes is often a parameter the econometrician wishes to conduct inference on. This is of empirical interest in in many fields such as labor, industrial organization and development. Furthermore, inference this type of system can be very complicated, if not impossible without strong parametric assumptions, which may not be reflected in the observed data. As we will discuss in detail, a semi parametric approach to these models, while desirable from a theoretical point of view because of its generality, often fail to achieve identification of parameter, or at best best only do so in sparse regions of the data, thus making inference impractical in practice. Given these two extremes- the non robustness of a parametric approach and the impracticality or even impossibility of conducting inference with semiparametric approaches in this setting, the factor structure condition may be a useful "in between" setting, and at the very least, can be used to gauge the sensitivity of the parametric approach to their stringent assumptions.

We will first illustrate our main idea within the context of a specific simultaneous model- the binary outcome with a binary endogenous explanatory variable which is modeled in a separate equation. We impose a factor structure to the two unobservables in this system and explore informational content of this assumption by comparing the identification results we attain to the extreme settings of fully parametric and fully semi parametric models. Our main findings in this case is that we no longer require the additional exclusion restriction nor the strong support conditions often assumed that was needed for identification in this model without the factor structure.

The rest of the paper is organized as follows. In the next section we formally describe the triangular system with factor structure, stating the conditions (e.g. regularity and support)

we impose on observed and unobserved variables. This is then followed by our main identification results for the parameters of interest in this model, notably the regression coefficients and the Average Structural Function(ASF). Many of these identification results translate directly into analogy based estimators of the parameters.

Section 3 explores the asymptotic properties of these estimators which fall into three categories. When the support conditions for point identification are not satisfied we establish set consistency of the estimator. When the support conditions result in irregular identification (identification at the boundary of support of observed variables) we show point consistency of the estimator. Finally, when the support conditions are such that regular identification is achieved, we show that the estimator is root-n consistent and asymptotically normal.

Section 4 explores the finite sample properties of the estimator in two ways. One is through simulation studies, and the other is an empirical illustration, where, following Dee (2004), we estimate the causal effect of civic returns to college.

Section 5 concludes with a summary and discussion for potential extensions of the base model, which involve the study of more complicated (non-triangular) systems, as well more complicated (nonparametric) factor structures. An appendix collects the proofs of the main theorems.

2 Triangular Model with Factor Structure

In this section we will consider the identification of the following factor structure model:

$$Y_1 = \mathbf{1}\{Z_1'\lambda_0 + Z_3'\beta_0 + \alpha_0 Y_2 - U > 0\}. \tag{2.1}$$

Turning to the model for the endogenous regressor, the binary endogenous variable Y_2 is assumed to be determined by the following reduced-form model:

$$Y_2 = \mathbf{1}\{Z'\delta_0 - V > 0\}, \tag{2.2}$$

where $Z \equiv (Z_1, Z_2)$ is the vector of "instruments" and (U, V) is a pair of random shocks. The subcomponent Z_2, Z_3 provides the exclusion restrictions in the model and is required to be nondegenerate conditional on $Z_1'\lambda_0 + Z_3'\beta_0$. We assume that the error terms U and V

are jointly independent of Z. The endogeneity of Y_2 in (2.1) arises when U and V are not independent.

The above system, or minor variations of it, have considered widely in the recent literature. See for example, Vytlacil and Yildiz (2007), Abrevaya, Hausman, and Khan (2010),Klein, Shan, and Vella (2011), Khan and Nekipelov (2010) and references therein.

An important parameter of interest is α_0, which relates to a treatment effect parameter. But as discussed in the aforementioned papers, this parameter is difficult, if not impossible to identify and estimate without imposing parametric restrictions on the unobserved variables in the model, (U, V). Such parametric restrictions, such as the often assumed bivariate normality assumption, are not robust to misspecification in the sense that any estimator of α_0 based on these conditions will be inconsistent if (U, V) have a different bivariate distribution.

The established difficulty of identifying α_0 in semi parametric, i.e., "distribution free" models, and the sensitivity of its identification to misspecification in parametric models is precisely what motivates the factor structure we add to the above model in this paper. Specifically, to allow for enodogeneity in the form of possible correlation between U, V, we augment the model add the following equation:

$$U = \gamma_0 V + \Pi \tag{2.3}$$

where Π is an unobserved random variable,, assumed to be distributed independently of (V, Z_1, Z_2, Z_3), and γ_0 is an additional unknown scalar parameter. This linear, one factor structure has been imposed in the literature many times- see for example Heckman (1991). Our goal will be to first establish identification for the parameters $(\alpha_0, \delta_0, \gamma_0, \beta_0, \lambda_0)$ under standard nonparametric regularity conditions on (U, V). [1] Later in the paper we will generalize the factor structure imposed here to consider nonlinear or nonparametric relationships between U, V. Our first results are based on the following conditions on both the observed variables $(Y_1, Y_2, Z_1, Z_2, Z_3)$ and unobserved variables (U, V, Π):, as well as parameter values $\alpha_0, \delta_0, \lambda_0, \beta_0, \gamma_0$.

So basically our approach is to add more structure to the fully semiparametric triangular binary system so quantify the identifying power of the added structure. Interestingly this is

[1] Actually, we will focus later in this paper on the parameters (α_0, γ_0). That is because the other parameter are not as difficult to identify, and work on them already exists in the literature.

the opposite approach of generalizing the fully parametric model. Such an approach has been taken recently in Han and Vytlacil (2013), who begin with a bivariate Probit model, and generalize it with the introduction of a class of one parameter Copulas, providing conditions such that identification can still be obtained. As we explain here, neither approach generalizes the other, as the two models are non nested.

The linear factor structure and the one-parameter copula model considered in Han and Vytlacil (2013) are not nested by each other. Based on the factor structure, we can recover F_Π, the distribution of Π, as a function of (F_U, F_V, λ) by deconvolution. Then we can write the copula of (U, V) as

$$F_{U,V}(F_U^{-1}(u), F_V^{-1}(v)) = \int_{-\infty}^{F_V^{-1}(v)} F_\Pi(F_U^{-1}(u) - \lambda w; F_U, F_V, \lambda) f_V(w) dw = C(u, v; F_U, F_V, \lambda).$$

If the marginals of (U, V) are known, then our linear factor structure implies the copula between (U, V) can be characterized by one parameter λ. However, comparing to Han and Vytlacil (2013), we do not require the copula to be stochastically increasing to achieve identification. If the marginals of (U, V) are unknown, then the copula depends not only on λ but only on two infinite dimensional parameter (F_U, F_V). Thus the factor structure cannot be characterized by a one-parameter copula. In addition, in order to achieve identification, Han and Vytlacil (2013) first nonparametrically identify the two marginals by assuming the existence of a full support common regressor. In contrast, under the factor structure, we bypass the nonparametric identification of the marginals as a whole and directly consider the identification of structure parameters. Therefore, in both cases, our model cannot be nested by the one-parameter copula model. On the other hand, there exists one-parameter copula models that cannot be decomposed into factor structures. This implies our model does not nest Han and Vytlacil (2013) either.

Our main identification results are based on the following conditions:

A1 The parameter $\theta_0 \equiv (\delta_0, \gamma_0, \beta_0, \lambda_0)$ is an element of a compact subset of R^4.

A2 The vector of unobserved variables, (U, V, Π) is continuously distributed with support on R^3 and independently distributed of the vector (Z_1, Z_2, Z_3). Furthermore, thaw random variable Π is distributed independently of V.

A3 The matrix $E[ZZ']$ is invertible, as is the matrix $E[\tilde{Z}\tilde{Z}']$ where $\tilde{Z} \equiv (Z_1, Z_3)$.

A4 The random variable Z_2 is continuously sitributed on an interval which is a subset of R, conditional on all values of $\tilde{Z}$.

A5 $|\alpha_0| < \ell(Z_1'\gamma_0 + Z_3'\beta_0) + \ell(Z_2)$, where $\ell(\cdot)$ denotes the length operator.

Under this set of conditions, we have the following identification result.

Theorem 2.1 *Under assumptions A1-A5, θ_0 is point identified.*

Proof : See appendix.

Thus the theorem concludes that under our stated conditions and our factor structure we can attain point identification. But what best demonstrates the identifying power of the factor structure is the comparison of our other assumptions compared to those typically imposed in the literature for this model. As explained in the remarks below the factor structure enables the relaxation of strong exclusion and support conditions typically assumed for inference in these types of models.

Remark 2.1 *Assumption A2 is standard in this literature in both the unobservables U, V as well as the independence between Π and V. References for the former (instruments independent of unobservables), can be found in Abrevaya, Hausman, and Khan (2010),Vytlacil and Yildiz (2007),Klein, Shan, and Vella (2011),Khan and Nekipelov (2010). For the latter, (Π independent of V), see, e.g. Chen, Khan, and Tang (2013),Chen and Khan (2008), Bai and Ng (2002).*

Remark 2.2 *Assumption A3 is the standard full rank condition found in these and other nonlinear models.*

Remark 2.3 *Assumption A4 requires the instrumental variable to be continuously distributed, which is often required in models with discrete outcomes. Recent papers - e.g. D'Haultfoeuille and Fevrier (2014), Torgovitsky (2014) establish identification with discrete instruments, but crucially require the endogenous variable in the outcome equation to be continuously distributed. This does not apply to our model nor many other treatment effect models.*

Remark 2.4 *Assumption A5 is in one sense a parameter space constraint. It is analogous to that imposed in Vytlacil and Yildiz (2007), but crucially distinct in important ways. Specifically, the length of the support of the instrument Z_2 now helps in the identification of α_0. This is natural, as a purpose of the instrument Z_2 should benefit in the identification of the parameters of the outcome equation as it sleds in standard IV approaches for the linear model. This is not the case, for example in Vytlacil and Yildiz (2007). Another crucial aspect of Assumption A5 is imposes no constraints on β_0. Specifically it can be 0, yet we still can attain identification. This is important to point out as without it, the econometrician would require the second exclusion restriction for identification, something difficult enough to attain in many empirical settings.*

Thus we immediately see informational content of the factor structure we impose. It enables point identification under weaker support condition when compared to the existing literature, and does not require the second exclusion restriction either. Later in the paper we will extend these arguments to the case where we do not attain point identification. Specifically, we will show that the factor structure enables sharper bounds for α_0 than bivariate models without factor structures, when point identification is not attainable in either model.

3 Estimation and Asymptotic Properties

The previous section established a point identification result, whose proof is given in the appendix. The identification result is constructive in the sense that it results directly in an analogy estimator for the parameters of interest which we describe in detail here. To simplify the exposition of our procedure, we will focus exclusively on the parameters α_0, γ_0. Thus we will treat the other parameters as known, and denote the resulting indexes, by X_1, X. Treating the other parameters as known can be justified by established results (see, e.g. Abrevaya, Hausman, and Khan (2010),Klein, Shan, and Vella (2011), Khan and Nekipelov (2010)) which show that these estimators are easier to identify and can be estimated at faster rates than α_0.

Denote $P^{ij}(x_1, x) = Prob(Y_1 = i, Y_2 = j | X_1 = x_1, X = x)$. These choice probabilities are unknown, but can be estimated as we describe below. Recall one of our identifying

assumptions was that the instrument in the treatment equation was continuously distributed. This assumption and our smoothness conditions on the choice probabilities implied that the choice probability functions were differentiable with respect to this instrument. Let $\partial_2 P^{ij}(x_1, x)$ denote the derivative of the ij choice probability with respect to the second argument, in this case the instrument in the treatment equation. This derivative is unknown, but is also estimable from the data.

Another function needed for our identification result was the density function of the unobserved term V, denoted by $f_V(\cdot)$. This is also unknown, but from the structure of our model can be recovered as the derivative (with respect to the instrument of $E[Y_1|X]$, and hence is estimable from the data. Our identification result depended on the sign of the index evaluated at two different regressor values:

$$X_1 + \alpha - \gamma X - (\tilde{X}_1 - \gamma \tilde{X})$$

where $(\tilde{X}_1, \tilde{X})$ denotes the second realization values.

As shown in the proof of identification, our main identification result

$$\partial_2 P^{11}(X_1, X)/f_V(X) + \partial_2 P^{10}(\tilde{X}_1, \tilde{X})/f_V(\tilde{X}) = 0 \quad \iff X_1 + \alpha - \gamma X - (\tilde{X}_1 - \gamma \tilde{X}) = 0$$

Note the left handed equality are functions of the data alone and not the unknown parameters. Furthermore, as said, while these functions, choice probability, density functions are unknown they can be consistently estimated from the data in a preliminary stage.

The right hand side equality involves the unknown parameters we wish to estimate and conduct inference on. As we will see, it will prove useful to rearrange the right handed equality as

$$\tilde{X}_1 - X_1 = \alpha + \gamma(\tilde{X} - X)$$

and note the above equator has a regression type form as if we were regression a "dependent" variable $\tilde{X}_1 - X_1$ on the "regressor" $(\tilde{X} - X)$, with "intercept" α and "slope coefficient" γ. This expression motivates a weighted least squares estimator of the unknown parameters α, γ, by only assigning positive weight to observations which satisfy the equality :

$$\partial_2 P^{11}(x_1, x)/f_V(x) + \partial_2 P^{10}(\tilde{x}_1, \tilde{x})/f_V(\tilde{x}) = 0$$

and for those observations regress $\tilde{X}_1 - X_1$ on $(\tilde{X} - X)$, with intercept.

Implementation requires further details to pay attention to, The unknown choice probabilities, their derivatives, and the density of V have to be estimated using nonparametric methods, and for this we adopt linear methods as they are particularly well suited for estimating derivatives of functions.

An additional implementation issue to deal with is that the equality above involving the choice probability derivatives will never occur exactly as they involve the instrumental variable which was assumed to be continuously distributed in our identification result. To address this problem, we sign "kernel" weights which depend on both how far the argument is from 0 (the further, the less weight) and the sample size, so in the limit we only use observations where the argument is arbitrarily close to 0. Such weights have been used in the literature in many settings - see e.g. Ahn and Powell (1993), Chen, Khan, and Tang (2013), for just a couple of many examples.

A last implementation issue we comment on here is the choice of $X, \tilde{X}$. Here we choose to use all pairs in the sample, which we denote by X_i, X_j. Thus from a sample of N observations our proposed estimator is to minimize the pairwise weighted least squares objective function:

$$\frac{1}{n(n-1)} \sum_{i \neq j} \hat{w}_{ij}((X_{1i} - X_{1j}) - \alpha - \gamma(X_i - X_j)^2) \tag{3.4}$$

where $\hat{w}_{ij}$ denotes the kernel weighting scheme.

We denote our estimator, which the minimizer of the the above objective function by $\hat{\alpha}, \hat{\gamma}$. In the next section we will discuss the asymptotic properties of this estimator under stated regularity conditions.

Before doing so, we discuss here both advantages and disadvantages of this estimation procedure:

Remark 3.1 *The proposed estimation procedure is computational friendly in that the sense that although it involves two stages, each stage is "closed form" in the sense that optimization routined are not required. One can simply do local linear estimation in the first stage to require derivatives of choice probabilities in the first stage, and weighted least squares in the second stage.*

Remark 3.2 *One disadvantage of the proposed procedure is the number of smoothing parameters required. Specifically, smoothing parameters are required in the first stage to estimate the derivatives of choice probabilities, and then again in the second stage to construct the kernel weights. The conditions on each of these tuning parameters, and how they relate to each other, are discussed in the next section, when we establish the asymptotic theory for this estimator.*

These advantages and disadvantages are worth further discussing when compared to a second procedure we introduce now. The first stage is identical and involves non parametrically estimating choice probability derivatives with respect to the continuous instrument. But the second stage involves a different objective function, which is more of the flavor of a least absolute deviations, as opposed to a least squares, approach: Letting $\hat{\theta}$ denote $(\hat{\alpha}, \hat{\gamma})$, our estimator is of the form:

The proposed estimator takes the following form:

$$\hat{\theta} = \arg\max_{\theta} Q_{n,2}(\theta) \equiv \sum_{i \neq j} \hat{g}_{i,j}(\theta)$$

in which

$$\begin{aligned} \hat{g}_{i,j}(\theta) &= [\mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} \\ &+ \mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\}], \end{aligned}$$

with

$$\phi(x_1, x, \tilde{x}_1, \tilde{x}; \theta) = x_1 + \alpha - \gamma x - (\tilde{x}_1 - \gamma \tilde{x})$$

This estimator has advantages and disadvantages when compared the first estimator. A main advantage is that only one smoothing parameter is required, in this case to estimate the derivative of the choice probability. However, a potential disadvantage is computational, as the second stage objective function is not smooth resulting in the estimator not being closed form. Optimization routines, such as Nelder- Meade for example, will be required for implementation.

The relative performance of each of the two estimators will be further explored later in the paper in the simulation studies section. As we will also see, under stated conditions, each of the two estimators is root-n consistent and asymptotically normal, so relative efficiency can be explored by comparing their asymptotic variances.

3.1 Average Structural Function

In the context of a single model such as the model with binary outcome and a binary endogenous regressor, there could be other parameters that are of interest. Thus far we have proposed a consistent estimator for the regression coefficient of a dummy endogenous variable in a triangular system with a binary outcome variable and showed how our factor structure could be useful in doing so. While the parameter α_0 is of interest, it is not the only parameter of interest from, say a policy perspective. For instance, a parameter of interest considered in Blundell and Powell (2004) is the *average structural function* (ASF). Generally speaking, the ASF is the predicted expected value of the outcome variable (Y) given the value of the explanatory variable (Y_1). For the binary model that we are considering in our example, identification of θ_0 and the uniform consistency of the estimator does not necessarily imply either of these properties for the ASF. The ASF can be expressed in terms of the marginal c.d.f. of the unobserved variable U (provided that $Y_2 = \mathbf{1}\{\theta_0 Y_1 + U > 0\}$) as $G(y_1) = 1 - F_U(\theta_0 y_1)$ (for $y_1 \in \{0, 1\}$).

where here $F_U(\cdot)$ denotes the cdf of U. So an estimator of $F_U(\cdot)$ as well as our estimator of α_0 would be required for inference on the ASF. (Blundell and Powell 2004) propose an estimator for the ASF in a binary outcome triangular system but require the endogenous variable to be continuously distributed which rules out the treatment effect model we wish to consider.

For the ASF here with a dummy endogenous variable, we propose the following multistep estimator:

1. Nonparametrically estimate (using, say, kernel methods) $E[D|Z]$ where recall D is the treatment variable and Z is the instrumental variable.
2. Construct "residuals", $\hat{V} = Y_1 - E[Y_1|Z]$.
3. With our estimator of α_0, denote by $\hat{\alpha}$, and our constructed residuals, $\hat{V}$, nonparametrically estimate, using, say, kernel methods, $E[Y_2|Y_1\hat{\theta}, \hat{V}]$.
4. To estimate the ASF at, say $Y_1 = 1$, integrate out with respect to $\hat{V}$:
   $\hat{G}(1) = \int \hat{E}[Y_1 \cdot \hat{\theta}, \hat{V}]d\hat{V}$

It remains to formally establish the asymptotic properties of this estimator, which will

be determined by our asymptotic properties developed for $\hat{\alpha}$. While in Blundell and Powell (2004) it was shown that the ASF is identified and can be consistently estimated when the endogenous variable Y_1 is continuous, such a result has not been established for the case where Y_1 is binary such as in our example. Moreover, since the distribution theory for $\hat{\alpha}$ is now standard, due to the regular identification enabled by the factor structure, we expect that the same will be true for the estimator of the ASF. Consequently, standard inference procedures will apply.

4 More General Factor Structures (Preliminary)

Up until now we have proposed identification and estimation results for a triangular system with a particular factor structure. As stated this particular factor structure was motivated y similar specifications previously imposed in the literature for different models. An advantage of our model specification was that it enabled stronger identification results for parameters of interest.

However a disadvantage of this structure was that it was restrictive in two ways. One is that it was a "one factor" model. The other is the linear in parameter relationship between the two unobserved components, which although often imposed in the literature, leaves open the possibility of misspecification. We leave the exploration to multi factor models to future research and focus in this section on a single, but nonlinear, nonparametric factor structure. As we can show here, the approach taken in the previous sections can readily extend to the more general model.

Specifically, we consider the following relationship between unobserved components:

$$U = g_0(V) + \tilde{\Pi} \tag{4.1}$$

where $\tilde{\Pi}$ is an unobserved random variable assumed to be distributed independently of V and all instruments. $g_0(\cdot)$ is an unknown function assumed to satisfy standard regularity conditions such as smoothness. Again, the parameter of interest is α, but now the unknown nuisance parameter in the factor equation is infinite dimensional. Now our approach is to replace the vector X with a series of basis functions of X, such as, for example orthonormal polynomials, in X. Those "basis" functions are meant to serve as an approximation of

$g_0(\cdot)$. With that replacement, we carry inexactly as before, except now instead of estimating a kernel weighted linear regression model it will be a kernel weighted semi linear, or partially linear regression model Robinson (1988). See the appendix for more details of how to construct such an estimator.

The asymptotic theory of this estimator for the generalized factor structure model will be based on the the number of basis functions increasing with the samples size, as is usually the case with series or sieve estimation, Ai and Chen (2003). Asymptotic properties of the estimator of the parameter α_0 can also be simultaneously recovered, as shown in Ai and Chen (2003).

5 Finite Sample Properties

In this section we explore the finite sample properties of the proposed estimation procedures via a simulation study. In both designs,

$$Y_1 = \mathbf{1}\{X_1 + \alpha_0 Y_2 - U \geq 0\},\ Y_2 = \mathbf{1}\{X - V > 0\}$$

such that (X_1, X) have marginals uniform (0.0.5) and $\mathcal{N}(0,1)$ respectively, are mutually independent and $(X_1, X) \perp (V, \Pi)$. $U = \gamma_0 V + \Pi$ such that (V, Π) are bivariate normal with zero mean and unit variance. For design 1, $(\alpha_0, \gamma_0) = (0.25, 0.5)$ such that both our and Vytlacil and Yildiz (2007)'s identification condition hold. For design 2, $(\alpha_0, \lambda_0) = (0.75, 0.5)$ such that our identification still holds while Vytlacil and Yildiz (2007)'s does not.

For each choice of sample size $n = 100, 200, 400$, we simulation 400 samples and report the bias, median bias, RMSE, MAD(median absolute deviation) of both Vytlacil and Yildiz (2007)'s estimator (WLS) and ours (WLS-F). For simplicity in implementation, we use second-order Gaussian kernels for matching the pairs and estimate the C.D.F. of (V, Π) and $\partial_2 P^{11}(x_1, x)/f_V(x) + \partial_2 P^{10}(\tilde{x}_1, \tilde{x})$ using local linear estimator, with bandwidth rates choses to minimize AMSE. Recent work by Henderson, Li, Parmeter, and Yao (2015) discuss bandwidth selection methods for estimating derivatives of regression functions which could prove useful for our estimator at hand, though we have yet to experiment with this.

We use bandwidth $h_1 = 0.7\hat{\sigma}_1 n^{-1/5}$ in the matching kernel in our estimator in which $\hat{\sigma}_1$ is the sample standard deviation of

$$\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j).$$

To compute Vytlacil and Yildiz (2007)'s estimator, we not only match $\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)$ with zero but also X_i with X_j. We use bandwidth $h_2 = 0.7\hat{\sigma}_2 n^{-1/5}$ and $h_x = 0.7\hat{\sigma}_x n^{-1/5}$ for the two match kernels respectively in which $\hat{\sigma}_2$ is the sample standard deviation of

$$\partial_2 P^{11}(X_{1,i}, X_i) + \partial_2 P^{10}(X_{1,j}, X_j)$$

and σ_x is $\sqrt{2}$ times the sample standard deviation of X. As results from the table indicate, the fiite sample performance generally agrees with the asymptotic theory. The estimator which foes not exploit the factor structure is clearly inconsistent for certain parameter values, as indicated by the bias not shrinking with the sample size. However, the RMSE for all estimators, including those that do impose the factor structure, do not appear to decline at the parametric rate. We attribute this to the rates chosen for the bandwidths, so clearly more work has to be done in this area.

$\alpha = 0.25$	Bias		Med. Bias		RMSE		MAD	
N	WLS-F	WLS	WLS-F	WLS	WLS-F	WLS	WLS-F	WLS
100	-0.007	-0.018	0.100	0.220	0.339	0.316	0.250	0.260
200	0.005	-0.055	0.120	0.240	0.331	0.289	0.240	0.260
400	0.119	-0.088	0.120	0.260	0.330	0.256	0.240	0.280
$\alpha = 0.75$	Bias		Med. Bias		RMSE		MAD	
N	WLS-F	WLS	WLS-F	WLS	WLS-F	WLS	WLS-F	WLS
100	-0.261	-0.277	0.360	0.240	0.361	0.288	0.400	0.260
200	-0.178	-0.254	0.340	0.240	0.351	0.259	0.380	0.240
400	-0.138	-0.245	0.340	0.240	0.341	0.248	0.360	0.240

Table 1: Finite sample performance

6 Application to the estimation of the civic returns to college (*in progress*)

We apply our method to the estimation of the civic returns to college, revisiting the influential work by Dee (2004) on this question. Specifically, we are interested in estimating the effect of attending college on several binary outcomes related to civic engagements and attitudes, including participation to votes and support for free speech. Our analysis is primarily based on data from the High School and Beyond longitudinal study, which follows over time a cohort of individuals who were high school sophomores in 1980. Follow-up interviews were conducted in 1984 and 1992. College attendance is reported in the 1984 interview, while the measures related to civic engagements and attitudes are obtained from the 1992 interview. Key to our empirical strategy is the availability of a continuous instrument for college attendance (our binary treatment here). We use the same instruments as Dee (2004), namely i) the distance from the respondent's high school to the nearest two-year college, and ii) the number of two-year colleges in the respondent's county in 1983. We refer the reader to Dee (2004) for a thorough discussion of the validity of these instruments in this context. Table 2 below reports the variables used in the analysis, along with their means. The total sample size is $N = 11,489$. From the Monte Carlo simulation results discussed above, we expect our estimator to perform well with this sample size, both in terms of bias and variance.

Unlike Dee (2004) who estimates the civic returns to college using a bivariate probit model, a key advantage of our method is that it is distribution-free.[2] It is worth noting that our framework is strictly more general than bivariate probit since our factor assumption is always satisfied when the joint distribution of the unobservables from the treatment and outcome equation is normal.

[2]See Altonji, Elder, and Taber (2005) who provide evidence on the role played by functional forms versus exclusion restrictions when using a bivariate probit to estimate the effect of catholic schooling on academic achievement.

Table 2: Variables from the High School and Beyond (Sophomore Cohort) data used in the Analysis

Variables	Mean
Currently registered to vote (1992)	0.669
Voted in past 12 months (1992)	0.355
Vote in 1988 Presidential election (1992)	0.553
Any volunteer work in last 12 months (1992)	0.371
High school graduate (1984)	0.844
College entrant (1984)	0.543
Importance of correcting inequality (1980)	1.8
Civics standardized test score (1980)	50.8
Female	0.521
Black	0.124
Hispanic	0.209
Other Race	0.051
Born Before 1964	0.284
Protestant	0.332
Catholic	0.382
Other Christian	0.047
Jewish	0.011
Other Religion	0.037
Religious background: none/missing	0.133
Family income missing	0.214
Family income <$8,000	0.06
Family income $8,000 to $14,999	0.117
Family income $15,000 to $19,999	0.105
Family income $20,000 to $24,999	0.109
Family income $25,000 to $29,999	0.106
Family income $30,000 to $39,999	0.127
Family income $40,000 to $49,999	0.071
Family income $50,000 or higher	0.092
Parent education missing	0.162
Parent high school dropout	0.282
Parent high school graduate	0.197
Parent some college	0.212
Parent college graduate	0.148
Single mother	0.136
Single father	0.027
Natural mother/stepfather	0.057
Natural father/stepmother	0.015
Other family structure	0.099
Both parents	0.666
School-level variables	
Urban school	0.227
Suburban school	0.503
Rural school	0.27
Miles to a 4-year college	16.7
Miles to a two-year college	16.7
State/county-level variables	
Number of two-year colleges in county	2.43
1980 county-level votes for President ÷ 18+ population	0.529
1980 county-level population aged 18 to 24	0.529
1980 county-level percent high school graduates among 25+ population	0.66
1992 state-level active mail-in voter registration	0.474
1992 state-level years with “motor-voter” regulations	1.4
Sample size	11,489

7 Conclusions

In this paper we explored the identifying power of factor structures in discrete simultaneous systems. We found that for a binary-binary system the factor structure we considered did indeed add informational content. Specifically, it enabled the relaxation of both the exclusion and support conditions typically employes in the identification of these models. As we then demonstrated factor structures then enabled the regular identification of parameters of interest, and we proposed new estimation procedures that converged at the parametric rate with a limiting normal distribution. Finite sample properties of the estimators were demonstrated thru simulation studies and an empirical illustration.

The work here opens ares for future research. For example, the factor structure we assume, while common in the existing literature, could be vernalized in different ways. For example, the structure could be more nonlinear an nonparametric. Although he outline a procedure for estimation in the latter case, a formal, more rigorous proof of the asymptotic theory for this procedure instill remains to be completed. Furthermore, models with multiple factors, and other nonlinear models are worth exploring. We leave these for future work.

References

Aakvik, A., J. Heckman, and E. J. Vytlacil (2005): "Estimating Treatment Effects for Discrete Outcomes when Responses to Treatment Vary: an Application to Norwegian Vocational Rehabilitation Programs," *Journal of Econometrics*, 125, 15–51.

Abbring, J., and J. Heckman (2007): "Econometrics Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in *Handbook of Econometrics, Vol. 6B*, ed. by J. J. Heckman, and E. E. Leamer. North Holland.

Abowd, J. M., and D. Card (1989): "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57, 411–445.

Abrevaya, J., J. Hausman, and S. Khan (2010): "Testing for Causal Effects in a Generalized Regression Model with Endogenous Regressors," *Econometrica*, 6, 2043–2061.

Ahn, H., and J. Powell (1993): "Semiparametric Estimation of Censored Selection Models," *Journal of Econometrics*, 58, 3–29.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1844.

Altonji, J., T. Elder, and C. Taber (2005): "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," *Journal of Human Resources*, 40, 791–821.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70(1), 191–221.

Blundell, R., and J. Powell (2004): "Endogeneity in Binary Response Models," *Review of Economic Studies*, 73.

Bonhomme, S., and J.-M. Robin (2010): "Generalized Non-Parametric Deconvolution with an Application to Earnings Dynamics," *Review of Economic Studies*, 77, 491–533.

Carneiro, P., K. Hansen, and J. J. Heckman (2003): "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44, 361–422.

Chen, S., and S. Khan (2008): "Rates of Convergence for Estimating Regression Coefficients in Heteroskedastic Discrete Response Models," *Journal of Econometrics*, 117, 245–278.

Chen, S., S. Khan, and X. Tang (2013): "On the Informational Content of Special Regressors in Heteroskedastic Binary Response Models," manuscript.

Cooley, J., S. Navarro, and Y. Takahashi (2015): "Identification and Estimation of Time-Varying Treatment Effects: How the Timing of Grade Retention Affects Outcomes," *Journal of Labor Economics*, forthcoming.

Cunha, F., and J. Heckman (2007): "Identifying and Estimating the Distributions of Ex Post and Ex Ante Returns to Schooling," *Labour Economics*, 14, 870–893.

Cunha, F., J. Heckman, and S. Navarro (2005): "Separating Uncertainty from Heterogeneity in Life Cycle Earnings," *Oxford Economic Papers*, 57, 191–261.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931.

Dee, T. S. (2004): "Are There Civic Returns to Education?," *Journal of Public Economics*, 88, 1697–1720.

D'Haultfoeuille, X., and P. Fevrier (2014): "Identification of Nonseparable Triangular Models with Discrete Instruments," *Econometrica*, forthcoming.

Donald, S., and W. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis*, 50, 30–40.

Han, S., and E. J. Vytlacil (2013): "Identification in a generalization of bivariate probit models with endogenous regressors," mimeograph, Yale University.

Heckman, J. (1991): "Identifying the Hand of the Past: Distinguishing State Dependence from Heterogeneity," *American Economic Review*, pp. 75–79.

Heckman, J., and S. Navarro (2007): "Dynamic Discrete Choice and Dynamic Treatment Effects," *Journal of Econometrics*, 136, 341–396.

Henderson, D., Q. Li, C. Parmeter, and S. Yao (2015): "Gradient-based Smootking Parameter Selection for Nonparametric Regression Estimation," *Journal of Econometrics*, 184, 233–241.

Horowitz, J. L., and M. Markatou (1996): "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63, 145–168.

Hull, J., and A. White (2004): "Valuation of a CDO and an nth to Default CDS without Monte Carlo Simulation," *Journal of Derivatives*, 12, 8–23.

Khan, S., and D. Nekipelov (2010): "Information Bounds and Impossibility Theorems for Simultaneous Discrete Response Models," Duke University Working Paper.

Khan, S., M. Ponomareva, and E. Tamer (2015): "Identification of Panel Data Models with Endogenous Censoring," Working Paper.

Khan, S., and E. Tamer (2010): "Irregular Identification, Support Conditions and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042.

Klein, R., C. Shan, and F. Vella (2011): "Semi-Parametric Estimation of Sample Selection Models with Binary Selection Rules and Binary Outcomes," Georgetown University working paper.

Manski, C. F. (1988): "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83.

Newey, W. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

Newey, W. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, pp. 233–253.

Oh, D., and A. Patton (2013): "Simulated Method of Moments Estimation for Copula-based Multivariate Models," *Journal of the American Statistical Association*, 108, 689–700.

——— (2015): "Modelling Dependence in High Dimensions with Factor Copulas," *Board of Governors of the FRB discussion series*, 2015-51.

Powell, J. (1994): "Estimation of Semiparametric Models," in *Handbook of Econometrics, Vol. IV.*, ed. by R. F. Engle, and D. McFadden, pp. 2444–2514. North-Holland.

Robinson, P. (1988): "Root-n-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Stock, J., and M. Watson (2011): "Dynamic Factor Models," *Handbook of Economic Forecasting*, 1, 35–59.

Torgovitsky, A. (2014): "Identification of Nonseparable Models Using Instruments with Small Support," *Econometrica*, forthcoming.

Vytlacil, E. J., and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75, 757–779.

A Identification Results for models with and without Factor Structrures

A.1 Without Factor Structure

Model: In tis section we will focus exclusively on the identification and estimation of the sub parameter vector $\tilde{\theta}_0 \equiv (\alpha_0, \lambda_0)$. We do so because the other parameters have already be shown to relatively easy to identify even without the factor structure- see, e.g. Abrevaya, Hausman, and Khan (2010). Without a factor structure, α_0 has been proven difficult to identify, see, e.g. Khan and Nekipelov (2010), and the identifiability λ_0 has remind an pen question in this literature. Thus we will focus on the "condensed" model:

$$Y_1 = \mathbf{1}\{X_1 + \alpha_0 Y_2 - U \geq 0\}$$
$$Y_2 = \mathbf{1}\{X - V \geq 0\}.$$

This section tries to derive the necessary and sufficient conditions for identification of α_0. What is the informational content of the model in the sense of Manski (1988). What is the semiparametric information bound in the sense of Newey (1990). When the identification assumption does not hold, what is the sharp identified set of α_0. Propose a so called adaptive estimation such that when identification is achieved, the object function is uniquely maximized at α_0 and when point-identification is not achieved, the object function is maximized at the identified set.

A.1.1 Conditions for identification

Assumption 1

1. $(X_1, X) \perp (U, V)$.

2. (X_1, X) *are continuously distributed with absolute continuous joint density w.r.t. Lebesgue measure. The support of* (X_1, X) *is* $[a, b] \times Supp(X)$*, in which* $Supp(X)$*, the support of* X*, is compact.*

3. V *is continuously distributed over* $\Re$*. And its density w.r.t. Lebesgue measure exist.*

Theorem A.1 *Assumption 1 holds. Then* $|\alpha_0| \leq b - a$ *is necessary and sufficient for* α_0 *to be identified.*

Proof : Denote $P^{ij}(x_1, x) = Prob(Y_1 = i, Y_2 = j | X_1 = x_1, X = x)$. Then

$$\begin{aligned} P^{11}(x_1, x) &= \int_{-\infty}^{x} F_U(x_1 + \alpha_0 | V = v) f(v) dv \\ P^{10}(\tilde{x}_1, x) &= \int_{x}^{+\infty} F_U(\tilde{x}_1 | V = v) f(v) dv. \end{aligned} \tag{A.1}$$

Taking derivatives w.r.t. the second argument of the the LHS function, we have

$$\begin{aligned} \partial_2 P^{11}(x_1, x) &= F_U(x_1 + \alpha_0 | V = x) f(x) \\ \partial_2 P^{10}(\tilde{x}_1, x) &= -F_U(\tilde{x}_1 | V = x) f(x). \end{aligned}$$

If $|\alpha_0| \leq b - a$, then there exists pair $(x_1, \tilde{x}_1)$ such that $x_1 + \alpha_0 = \tilde{x}_1$. This pair can be identified by checking the equation below:

$$\partial_2 P^{11}(x_1, x)/f(x) + \partial_2 P^{10}(\tilde{x}_1, x)/f(x) = 0.$$

This concludes the sufficient part.

When $\alpha_0 < a - b$, for any $\alpha < \alpha_0$, we can define

$$\begin{aligned} &\tilde{U} = U + \alpha - \alpha_0 && \text{if} && U \leq b + \alpha_0 \\ &\tilde{U} = U && \text{if} && U > b + \alpha_0 \end{aligned}$$

Then for any $x_1 \in [a, b]$,

$$\begin{aligned} P(\tilde{U} \leq x_1 + \alpha | V = v) &= P(\tilde{U} \leq x_1 + \alpha, U \leq b + \alpha_0) + P(\tilde{U} \leq x_1 + \alpha, U > b + \alpha_0 | V = v) \\ &= P(U \leq x_1 + \alpha_0 | V = v) \\ P(\tilde{U} \leq x_1 | V = v) &= P(\tilde{U} \leq x_1, U \leq b + \alpha_0 | V = v) + P(\tilde{U} \leq x_1, U > b + \alpha_0 | V = v) \\ &= P(U \leq b + \alpha_0, U \leq x_1 + \alpha_0 - \alpha | V = v) + P(b + \alpha_0 < U \leq x_1, | V = v) \\ &= P(U \leq b + \alpha_0 | V = v) + P(b + \alpha_0 < U \leq x_1, | V = v) \\ &= P(U \leq x_1 | V = v). \end{aligned}$$

Let $G_{U,V}$ and $G_{\tilde{U},V}$ be the joint distribution of (U, V) and $(\tilde{U}, V)$ respectively. Then the above calculation with (A.1) imply that $(\alpha_0, G_{U,V})$ and $(\alpha, G_{\tilde{U},V})$ are observationally equivalent.

When $\alpha_0 > b - a$, for any $\alpha > \alpha_0$, we can define

$$\begin{aligned} &\tilde{U} = U + \alpha - \alpha_0 && \text{if} && U > a + \alpha_0 \\ &\tilde{U} = U && \text{if} && U \leq a + \alpha_0 \end{aligned}$$

Then for any $x_1 \in [a, b]$,

$$\begin{aligned} P(\tilde{U} \leq x_1 + \alpha | V = v) &= P(\tilde{U} \leq x_1 + \alpha, U \leq a + \alpha_0) + P(\tilde{U} \leq x_1 + \alpha, U > a + \alpha_0 | V = v) \\ &= P(U \leq a + \alpha_0 | V = v) + P(a + \alpha_0 < U \leq x_1 + \alpha_0 | V = v) \\ &= P(U \leq x_1 + \alpha_0 | V = v). \\ P(\tilde{U} \leq x_1 | V = v) &= P(\tilde{U} \leq x_1, U \leq a + \alpha_0 | V = v) + P(\tilde{U} \leq x_1, U > a + \alpha_0 | V = v) \\ &= P(U \leq x_1 | V = v). \end{aligned}$$

So again, $(\alpha_0, G_{U,V})$ and $(\alpha, G_{\tilde{U},V})$ are observationally equivalent.

Theorem A.2 *Assumption 1 holds. When $|\alpha_0| > b - a$, the sharp identified set for α_0 is*

$$\mathcal{A}^* = \{\alpha : \alpha > b - a \text{ if } \alpha_0 > 0 \text{ and } \alpha < a - b \text{ if } \alpha_0 < 0\}.$$

Proof : First, when $|\alpha_0| > b - a$, the sign of α_0 is identified by the data. We take $\alpha_0 > b - a$ as an example. By the proof of Theorem A.1, we have already shown that all $\alpha > \alpha_0$ is in the identified set. Now we consider $\frac{b-a+\alpha_0}{2} \leq \alpha < \alpha_0$.

$$\begin{aligned} &\tilde{U} = U + \alpha - \alpha_0 && \text{if} && U > a + \alpha \\ &\tilde{U} = U && \text{if} && U \leq a + \alpha \end{aligned}$$

Then for any $x_1 \in [a, b]$,

$$\begin{aligned} P(\tilde{U} \leq x_1 + \alpha | V = v) &= P(\tilde{U} \leq x_1 + \alpha, U \leq a + \alpha) + P(\tilde{U} \leq x_1 + \alpha, U > a + \alpha | V = v) \\ &= P(U \leq a + \alpha | V = v) + P(a + \alpha < U \leq x_1 + \alpha_0 | V = v) \\ &= P(U \leq x_1 + \alpha_0 | V = v). \\ P(\tilde{U} \leq x_1 | V = v) &= P(\tilde{U} \leq x_1, U \leq a + \alpha | V = v) + P(\tilde{U} \leq x_1, U > a + \alpha | V = v) \\ &= P(U \leq x_1 | V = v) + P(U \leq x_1 + \alpha_0 - \alpha, U > a + \alpha | V = v). \\ &= P(U \leq x_1 | V = v). \end{aligned}$$

Here note that the last equality is because $x_1 + \alpha_0 - \alpha \leq b + \alpha_0 - \alpha \leq a + \alpha$ if $\alpha \geq \frac{b-a+\alpha_0}{2}$. Denote $\alpha^{(1)} = \frac{b-a+\alpha_0}{2}$. Then we have shown that there exists $U^{(1)}(\alpha)$ which only depends on α such that for any $x1 \in [a, b]$, any v and any $\alpha_0 > \alpha \geq \alpha^{(1)}$

$$\begin{aligned} &P(U^{(1)}(\alpha) \leq x_1 + \alpha | V = v) = P(U \leq x_1 + \alpha_0 | V = v) \\ &P(U^{(1)}(\alpha) \leq x_1 | V = v) = P(U \leq x_1 | V = v). \end{aligned}$$

In particular, there exists $U^{(1)}(\alpha^{(1)})$ such that

$$P(U^{(1)}(\alpha^{(1)}) \leq x_1 + \alpha^{(1)}|V = v) = P(U \leq x_1 + \alpha_0|V = v)$$
$$P(U^{(1)}(\alpha^{(1)}) \leq x_1|V = v) = P(U \leq x_1|V = v).$$

Now repeating the above construction but replacing U with $U^{(1)}$ and α_0 with $\alpha^{(1)}$, we have for any $\alpha^{(1)} > \alpha \geq \alpha^{(2)} \equiv \frac{b-a+\alpha^{(1)}}{2}$, there exists $U^{(2)}(\alpha)$ such that for any $x_1 \in [a, b]$, any v and any $\alpha^{(1)} > \alpha \geq \alpha^{(2)}$,

$$P(U^{(2)}(\alpha) \leq x_1 + \alpha^{(2)}|V = v) = P(U^{(1)}(\alpha^{(1)}) \leq x_1 + \alpha^{(1)}|V = v) = P(U \leq x_1 + \alpha_0|V = v)$$
$$P(U^{(2)}(\alpha) \leq x_1|V = v) = P(U^{(1)}(\alpha^{(1)}) \leq x_1|V = v) = P(U \leq x_1|V = v).$$

This concludes that any α such that $\alpha_0 > \alpha \geq \alpha^{(2)}$ is in the identified set. In general, by repeating the procedure k times, we have that any α such that

$$\alpha_0 > \alpha \geq \alpha^{(k)} = (1 - \frac{1}{2^k})(b - a) + \frac{\alpha_0}{2^k}$$

is in the identified set. For any $\alpha > b - a$, there exists some finite k such that $\alpha > (1 - \frac{1}{2^k})(b - a) + \frac{\alpha_0}{2^k}$. This concludes the result that $\alpha > b - a$ is in the identified set.

Finally, since if $\alpha > b-a$, $\partial_2 P^{11}(x_1, x)+\partial_2 P^{10}(\tilde{x}_1, x) > 0$ for all pairs of (x_1, x) and $(\tilde{x}_1, x)$ while, if $\alpha \leq b - a$, at least there exists one pair (x_1, x) and $(\tilde{x}_1, x)$ such that $\partial_2 P^{11}(x_1, x) + \partial_2 P^{10}(\tilde{x}_1, x) \leq 0$. This implies $\alpha \leq b - a$ is not in the identified set. Therefore, the sharp identified set when $\alpha_0 > b - a$ is $\alpha > b - a$.

When $\alpha_0 < a - b$, symmetric argument implies that the identified set is $\alpha < a - b$.

A.1.2 Adaptive Estimator

Now we propose an estimator of α_0. Based on the proof of theorem A.1, if $\alpha_0 \leq b - a$,

$$\{\partial_2 P^{11}(X_1, X) + \partial_2 P^{10}(\tilde{X}_1, \tilde{X}) \geq 0\}\mathbf{1}\{X = \tilde{X}\} \iff \{X_1 + \alpha_0 \geq \tilde{X}_1\}\mathbf{1}\{X = \tilde{X}\}.$$

We can nonparametrically estimate the LHS, so the sample object function is

$$\hat{\alpha} = \arg\max \sum_{i \neq j} \hat{f}_{i,j}(\alpha)$$

in which

$$\begin{aligned}\hat{f}_{i,j}(\alpha) =& k(\frac{X_i - X_j}{h})[\mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j) \geq 0\}\mathbf{1}\{X_{1,i} + \alpha \geq X_{1,j}\} \\ &+ \mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j) < 0\}\mathbf{1}\{X_{1,i} + \alpha < X_{1,j}\}].\end{aligned}$$

It is easy to see that the infeasible object function take the form of $Q(\alpha) = E(f_{i,j}(\alpha)|X_i = X_j)$ in which

$$\begin{aligned} f_{i,j}(\alpha) =&[\mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i) + \partial_2 P^{10}(X_{1,j}, X_i) \geq 0\}\mathbf{1}\{X_{1,i} + \alpha \geq X_{1,j}\} \\ &+ \mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i) + \partial_2 P^{10}(X_{1,j}, X_i) < 0\}\mathbf{1}\{X_{1,i} + \alpha < X_{1,j}\}]. \end{aligned}$$

Theorem A.3 *Assumption 1 holds. If* $(X_1, \tilde{X}_1)|X = \tilde{X}$, $|\alpha_0| < b - a$, *then* α_0 *is the unique maximizer of* $Q(\alpha)$.

Proof : For any $\alpha \neq \alpha_0$,

$$\begin{aligned} Q(\alpha_0) - Q(\alpha) =&E((\mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i) + \partial_2 P^{10}(X_{1,j}, X_i) \geq 0\} - \mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i) + \partial_2 P^{10}(X_{1,j}, X_i) < 0\}) \\ &\times [\mathbf{1}\{X_{1,i} + \alpha_0 \geq X_{1,j} > X_{1,i} + \alpha\} - \mathbf{1}\{X_{1,i} + \alpha_0 < X_{1,j} \leq X_{1,i} + \alpha\}]|X_i = X_j) \end{aligned} \tag{A.2}$$

So if $\alpha > \alpha_0$,

$$Q(\alpha_0) - Q(\alpha) = P(X_{1,i} + \alpha_0 \leq X_{1,j} < X_{1,i} + \alpha).$$

Note that the support of $(X_{1,i}, X_{1,j})|X_i = X_j$ is $[a, b] \times [a, b]$. The area $x_{1,i} + \alpha_0 \leq x_{1,j} < x_{1,i} + \alpha, (x_{1,i}, x_{1,j}) \in [a, b] \times [a, b]$ has a positive Lebesgue measure as shown in the Figure below. Then since $(X_1, \tilde{X}_1)|X = \tilde{X}$ is absolute continuous, $P(X_{1,i} + \alpha_0 \leq X_{1,j} < X_{1,i} + \alpha|X_i = X_j) > 0$.

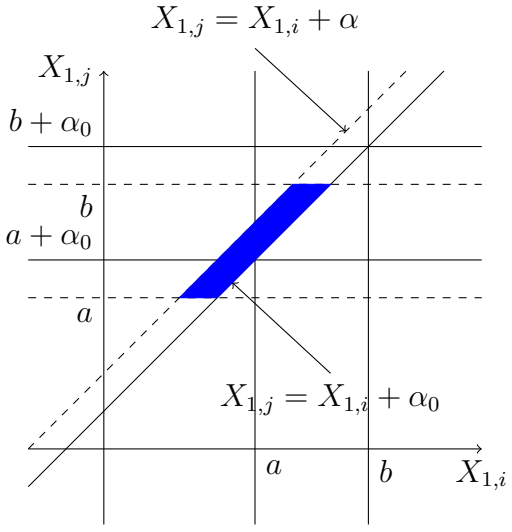

Figure 1: Positive Probability when $0 < \alpha_0 < b - a$ and $\alpha > \alpha_0$

Similarly, when $\alpha < \alpha_0$,

$$Q(\alpha_0) - Q(\alpha) = P(X_{1,i} + \alpha_0 \geq X_{1,j} > X_{1,i} + \alpha | X_i = X_j) > 0.$$

A.1.3 Manski's information content

$\mathcal{A}_1(\alpha) = \{(X_1, \tilde{X}_1), \phi(X_1, \tilde{X}_1; \alpha_0) \geq 0 > \phi(X_1, \tilde{X}_1; \alpha)$ or $\phi(X_1, \tilde{X}_1; \alpha_0) \leq 0 < \phi(X_1, \tilde{X}_1; \alpha)\}$, in which

$$\phi(x_1, \tilde{x}_1; \alpha) = x_1 + \alpha - \tilde{x}_1.$$

In Theorem A.3, we have shown that when $|\alpha_0| < b - a$, $P(\mathcal{A}(\alpha)) > 0$ for any $\alpha \neq \alpha_0$. Next we consider the case when $|\alpha_0| \geq b - a$. Denote $\overline{\mathcal{A}}_1 = \{\alpha : P(\mathcal{A}_1(\alpha)|X = \tilde{X}) = 0\}$. Then we call $\overline{\mathcal{A}}_1$ that we cannot distinguish from the true parameter α_0.

Theorem A.4 *Assumption 1 holds.* $(X_1, \tilde{X}_1)|X = \tilde{X}$. *Then if* $|\alpha_0| > b - a$,

$$\overline{\mathcal{A}}_1 = \{\alpha : \alpha \geq b - a \text{ if } \alpha_0 > 0 \text{ and } \alpha \leq a - b \text{ if } \alpha_0 < 0\}.$$

Proof : First, as in the proof of Theorem A.2, we note that when $|\alpha_0| > b - a$, the sign of α_0 is identified. For $\alpha_0 > 0$, it is easy to see that if $\alpha_0 > b - a$, $x_1 + \alpha_0 > \tilde{x}_1$ for any $(x_1, \tilde{x}_1) \in \text{Supp}(X_1) \times \text{Supp}(X_1)$. For any $\alpha \geq b - a$, $P(\mathcal{A}_1(\alpha)|X = \tilde{X}) = 0$ and for any $\alpha < b - a$, $\lambda(\mathcal{A}_1(\alpha)) > 0$. Since $(X_1, \tilde{X}_1)|X = \tilde{X}$ is absolutely continuous, $P(\mathcal{A}_1(\alpha)|X = \tilde{X}) > 0$. The case for $\alpha_0 < 0$ can be proved symmetrically. This concludes the proof.

Comparing Manski's information content $\overline{A}_1$ with the sharp identified set $\mathcal{A}^*$ in Theorem A.2, we notice that the only different is the boundary scenario that $|\alpha_0| = b - a$. In this case, only two boundary points $x_1 = a, \tilde{x}_1 = b$ are useful for identifying α_0. Since X_1 has an absolute continuous distribution on its support $[a, b]$, identification is achieved at a zero-probability set. By Khan and Tamer (2010), we call this identification irregular. Theorem A.4 shows that the adaptive estimation cannot distinguish irregular identification from non-identification. In fact, next section shows that when $|\alpha_0| < b - a$, the semiparametric efficiency bound for α_0 is positive, while when $|\alpha_0| = b - a$, the semiparametric efficiency bound becomes zero.

Theorem A.5 *Assumption 1 holds. If $|\alpha_0| = b - a$, then the semiparametric efficiency bound for α_0 is zero.*

Proof: Recall that $P^{ij}(x_1, x) = P(Y_1 = i, Y_2 = j | X_1 = x_1, X = x)$. We also denote $F(t_1, t_2) = Prob(U \leq t_1, V \leq t_2)$. Then we have

$$
\begin{aligned}
&P^{11}(x_1, x) = F(x_1 + \alpha_0, x), \\
&P^{10}(x_1, x) = F(x_1, \infty) - F(x_1, x), \\
&P^{01}(x_1, x) = F(\infty, x) - F(x_1 + \alpha_0, x), \\
&P^{00}(x_1, x) = 1 - F(x_1, \infty) - F(\infty, x) + F(x_1, x).
\end{aligned}
$$

Because F is the only infinite dimensional parameter. We now consider a one parameter submodel of F which is written as $\lambda(t_1, t_2) = \delta h(t_1, t_2) + F(t_1, t_2)$. For a model for some fixed δ, we denote $P_\delta^{ij}(x_1, x) = P_\delta(Y_1 = i, Y_2 = j | X_1 = x_1, X = x)$, such that

$$
\begin{aligned}
&P_\delta^{11}(x_1, x) = \lambda(x_1 + \alpha_0, x), \\
&P_\delta^{10}(x_1, x) = \lambda(x_1, \infty) - \lambda(x_1, x), \\
&P_\delta^{01}(x_1, x) = \lambda(\infty, x) - F(x_1 + \alpha_0, x), \\
&P_\delta^{00}(x_1, x) = 1 - \lambda(x_1, \infty) - \lambda(\infty, x) + \lambda(x_1, x).
\end{aligned}
$$

Let $G(t_1, t_2) = \frac{\partial F(t_1, t_2)}{\partial t_1}$. Then the score of α_0 is

$$
\Psi_\alpha(x_1, x) = \frac{1}{2}[\frac{y_1 y_2}{F^{1/2}(x_1 + \alpha_0, x)} - \frac{(1 - y_1) y_2}{(F(\infty, x) - F(x_1 + \alpha_0, x))^{1/2}}] G(x_1 + \alpha_0, x).
$$

Similarly, the score for F is

$$
\begin{aligned}
\Psi_\delta(x_1, x) =& \frac{1}{2}[\frac{y_1 y_2 h(x_1 + \alpha_0, x)}{F^{1/2}(x_1 + \alpha_0, x)} + \frac{(1 - y_1) y_2 (h(\infty, x) - h(x_1 + \alpha_0, x))}{(F(\infty, x) - F(x_1 + \alpha_0, x))^{1/2}} \\
&+ \frac{y_1 (1 - y_2)(h(x_1, \infty) - h(x_1, x))}{(F(x_1, \infty) - F(x_1, x))^{1/2}} \\
&+ \frac{(1 - y_1)(1 - y_2)(h(x_1, x) - h(x_1, \infty) - h(\infty, x))}{(1 - F(x_1, \infty) - F(\infty, x) + F(x_1, x))^{1/2}}].
\end{aligned}
$$

The information for the one-parameter family is

$$
\begin{aligned}
&\int 4(\Psi_\alpha - \Psi_\delta)^2 d\mu \\
=& \int (\frac{(G(x_1 + \alpha_0, x) - h(x_1 + \alpha_0, x))^2}{P^{11}(x_1, x)} + \frac{(h(x_1, \infty) - h(x_1, x))^2}{P^{10}(x_1, x)} \\
&+ \frac{(G(x_1 + \alpha_0, x) + h(\infty, x) - h(x_1 + \alpha_0, x))^2}{P^{01}(x_1, x)} + \frac{(h(x_1, \infty) + h(\infty, x) - h(x_1, x))^2}{P^{00}(x_1, x)}) dF(x_1, x)
\end{aligned}
$$

We now consider the case in which $\alpha_0 > 0$ and $\alpha_0 = b - a$. Since $\text{Supp}(X)$ is compact, we can let $h(x_1, \infty) = h(\infty, x) = 0$. The we can choose $h(t_1, t_2) = G(t_1, t_2)$ for $(t_1, t_2) \in [b, b + \alpha_0] \times \text{Supp}(X)$ and $h(t_1, t_2) = 0$ for $(t_1, t_2) \in [a, b-\eta] \times \text{Supp}(X)$. On $(t_1, t_2) \in [b-\eta, b] \times \text{Supp}(X)$, $h(t_1, t_2) = \frac{t_1 - b + \eta}{\eta} G(b, x)$. Then we have

$$\begin{aligned}
&\int 4(\Psi_\alpha - \Psi_\delta)^2 d\mu \\
&= \int (\frac{x_1 - b + \eta}{\delta})^2 \mathbf{1}\{b - \eta \leq x_1 \leq b\} [\frac{G^2(b, x)}{P^{10}(x_1, x)} + \frac{G^2(b, x)}{P^{00}(x_1, x)}] dF(x_1, x).
\end{aligned}$$

By letting $\eta w = b - x_1$, we have

$$\begin{aligned}
&\int 4(\Psi_\alpha - \Psi_\delta)^2 d\mu \\
&= \int (\frac{x_1 - b + \eta}{\eta})^2 \mathbf{1}\{b - \eta \leq x_1 \leq b\} [\frac{G^2(b, x)}{P^{10}(x_1, x)} + \frac{G^2(b, x)}{P^{00}(x_1, x)}] dF(x_1, x) \\
&\lesssim \int_0^1 (1 - w)^2 \eta f(b - w\eta, x) dw dx \to 0
\end{aligned}$$

as $\eta \to 0$. The case in which $\alpha_0 = a - b$ can be proved by the same manner.

A.2 With Factor Structure

Model:

$$\begin{aligned}
Y_1 &= \mathbf{1}\{X_1 + \alpha_0 Y_2 - U \geq 0\} \\
Y_2 &= \mathbf{1}\{X - V \geq 0\},
\end{aligned}$$

where $U = \gamma_0 V + \Pi$ and $V \perp \Pi$. In this section, we want to propose another adaptive estimation procedure and consider the information content explored by that. We then compare the two information contents and argue that the one with factor structure is strictly large than the one without. This implies two scenarios. (1) Factor structure helps identifying α_0 when it is not without it. (2) In both case, α_0 is not identified. But the adaptive estimation produce narrower identified set when imposing factor structure. Note here we are not sure the new adaptive estimation explore all the information content, i.e. the new identified set when imposing factor structure is not necessarily sharp.

A.2.1 Conditions for identification

Assumption 2

1. $(X_1, X) \perp (U, V)$.
2. (X_1, X) *are continuously distributed with absolute continuous joint density w.r.t. Lebesgue measure with compact support. The density is bounded and bounded away from zero on the support.*
3. V *is continuously distributed over* $\Re$*. And its density w.r.t. Lebesgue measure exist.*

The identification relies on overlap support and a rank condition. The overlap support condition is similar to the one for model without factor structure. But it also takes into account of the variation of X. The rank condition is new because here we have two unknown parameters. Rank condition helps to identify them separately from a system of equations.

Assumption 3

1. $Supp(X_1 + \alpha_0 - \gamma_0 X) \cap Supp(X_1 - \gamma_0 X) \neq \emptyset$.
2. *For any constant* c, $P(X - \tilde{X} = c | X_1 + \alpha_0 - \gamma_0 X = \tilde{X}_1 - \gamma_0 \tilde{X}) < 1$.

Theorem A.6 *Assumption 2 and 3 hold. Then* $\theta_0 \equiv (\alpha_0, \gamma_0)$ *is identified.*

Proof:

$$P^{11}(x_1, x) = \int_{-\infty}^{x} F_\Pi(x_1 + \alpha_0 - \gamma_0 v) f_V(v) dv$$
$$P^{10}(\tilde{x}_1, \tilde{x}) = \int_{\tilde{x}}^{+\infty} F_\Pi(\tilde{x}_1 - \gamma_0 v) f_V(v) dv.$$

Taking derivatives w.r.t. the second argument of the LHS function, we obtain

$$\partial_2 P^{11}(x_1, x) / f_V(x) = F_\Pi(x_1 + \alpha_0 - \gamma_0 x)$$
$$-\partial_2 P^{10}(\tilde{x}_1, \tilde{x}) / f_V(\tilde{x}) = F_\Pi(\tilde{x}_1 - \gamma_0 \tilde{x}).$$

By Assumption 3-1, we know that there exists pairs $(x_1^{(1)}, x^{(1)})$ and $(\tilde{x}_1^{(1)}, \tilde{x}^{(1)})$ in $\text{Supp}(X_1, X)$ such that

$$x_1^{(1)} + \alpha_0 - \gamma_0 x^{(1)} = \tilde{x}_1^{(1)} - \gamma_0 \tilde{x}^{(1)}.$$

These pairs can be identified from data by the fact that

$$\partial_2 P^{11}(x_1^{(1)}, x^{(1)})/f_V(x^{(1)}) + \partial_2 P^{10}(\tilde{x}_1^{(1)}, \tilde{x}^{(1)})/f_V(\tilde{x}^{(1)}) = 0.$$

By Assumption 3-2, there exists at least another pair $(x_1^{(2)}, x^{(2)})$ and $(\tilde{x}_1^{(2)}, \tilde{x}^{(2)})$ in $\text{Supp}(X_1, X)$ such that

$$x_1^{(2)} + \alpha_0 - \gamma_0 x^{(2)} = \tilde{x}_1^{(2)} - \gamma_0 \tilde{x}^{(2)}, \text{ and } x^{(2)} - \tilde{x}^{(2)} \neq x^{(1)} - \tilde{x}^{(1)}.$$

So we have a two equation system

$$\begin{aligned}
\alpha_0 - \gamma_0(x^{(1)} - \tilde{x}^{(1)}) &= \tilde{x}_1^{(1)} - x_1^{(1)} \\
\alpha_0 - \gamma_0(x^{(2)} - \tilde{x}^{(2)}) &= \tilde{x}_1^{(2)} - x_1^{(2)}.
\end{aligned}$$

Since $x^{(2)} - \tilde{x}^{(2)} \neq x^{(1)} - \tilde{x}^{(1)}$, the system of equations has a unique solution. This concludes the proof.

A.2.2 Adaptive Estimator

Recall that we have

$$\begin{aligned}
\partial_2 P^{11}(x_1, x)/f_V(x) &= F_\Pi(x_1 + \alpha_0 - \gamma_0 x) \\
-\partial_2 P^{10}(\tilde{x}_1, \tilde{x})/f_V(\tilde{x}) &= F_\Pi(\tilde{x}_1 - \gamma_0 \tilde{x}).
\end{aligned}$$

The proposed estimator takes the following form:

$$\hat{\theta} = \arg\max_\theta Q_{n,2}(\theta) \equiv \sum_{i \neq j} \hat{g}_{i,j}(\theta)$$

in which

$$\begin{aligned}
\hat{g}_{i,j}(\theta) =& [\mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} \\
&+ \mathbf{1}\{\partial_2 \hat{P}^{11}(X_{1,i}, X_i)/\hat{f}_V(X_i) + \partial_2 \hat{P}^{10}(X_{1,j}, X_j)/\hat{f}_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\}],
\end{aligned}$$

and

$$\phi(x_1, x, \tilde{x}_1, \tilde{x}; \theta) = x_1 + \alpha - \gamma x - (\tilde{x}_1 - \gamma \tilde{x}).$$

The infeasible kernel for the U-statistic is

$$\begin{aligned} g_{i,j}(\theta) =& \mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j) \geq 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) \geq 0\} \\ &+ \mathbf{1}\{\partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j) < 0\}\mathbf{1}\{\Phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta) < 0\}. \end{aligned}$$

Denote $Q_2(\theta) = Eg_{i,j}(\theta)$.

Assumption 4 *There exists a open set $S \subset Supp(X_1, X) \times Supp(X_1, X) \subset \Re^4$ such that $S \cap \{(x_1, x, \tilde{x}_1, \tilde{x}) : \phi(x_1, x, \tilde{x}_1, \tilde{x}; \theta_0) = 0\} \neq \emptyset$.*

Theorem A.7 *Assumption 2, 4 hold, then $Q_2(\theta)$ to have a unique maximizer.*

Proof: Denote $G_{i,j} = \partial_2 P^{11}(X_{1,i}, X_i)/f_V(X_i) + \partial_2 P^{10}(X_{1,j}, X_j)/f_V(X_j)$ and for simplicity, $\phi_{i,j}(\theta) = \phi(X_{1,i}, X_i, X_{1,j}, X_j; \theta)$, then

$$\begin{aligned} & Q_2(\theta_0) - Q_2(\theta) \\ =& E(\mathbf{1}\{G_{i,j} \geq 0\} - \mathbf{1}\{G_{i,j} < 0\})[\mathbf{1}\{\phi_{i,j}(\theta_0) \geq 0 > \phi_{i,j}(\theta)\} - \mathbf{1}\{\phi_{i,j}(\theta) \geq 0 > \phi_{i,j}(\theta_0)\}] \\ =& P(\phi_{i,j}(\theta_0) \geq 0 > \phi_{i,j}(\theta)) + P(\phi_{i,j}(\theta) \geq 0 > \phi_{i,j}(\theta_0)). \end{aligned}$$

Hence, θ_0 is the unique solution if and only if for any $\theta \neq \theta_0$,

$$P(\phi_{i,j}(\theta_0) \geq 0 > \phi_{i,j}(\theta)) + P(\phi_{i,j}(\theta) \geq 0 > \phi_{i,j}(\theta_0)) > 0.$$

Since $S \cap \{(x_1, x, \tilde{x}_1, \tilde{x}) : \phi(x_1, x, \tilde{x}_1, \tilde{x}; \theta - 0) = 0\} \neq \emptyset$, S is open and (X_1, X) is absolutely continuous w.r.t. Lebesgue measure, $P(\phi_{i,j}(\theta_0) \geq 0, (X_{1,i}, X_i, X_{1,j}, X_j) \in S) > 0$ and $P(\phi_{i,j}(\theta_0) < 0, (X_{1,i}, X_i, X_{1,j}, X_j) \in S) > 0$. Since $\theta \neq \theta_0$, at least one of $\{\phi_{i,j}(\theta_0) \geq 0 > \phi_{i,j}(\theta), (X_{1,i}, X_i, X_{1,j}, X_j) \in S\}$ or $\{\phi_{i,j}(\theta_0) < 0 \leq \phi_{i,j}(\theta), (X_{1,i}, X_i, X_{1,j}, X_j) \in S\}$ is nonempty. This implies

$$\begin{aligned} & P(\phi_{i,j}(\theta_0) \geq 0 > \phi_{i,j}(\theta)) + P(\phi_{i,j}(\theta) \geq 0 > \phi_{i,j}(\theta_0)) \\ \geq& P(\phi_{i,j}(\theta_0) \geq 0 > \phi_{i,j}(\theta), (X_{1,i}, X_i, X_{1,j}, X_j) \in S) + P(\phi_{i,j}(\theta) \geq 0 > \phi_{i,j}(\theta_0), (X_{1,i}, X_i, X_{1,j}, X_j) \in S) > 0. \end{aligned}$$

A.2.3 Manski's Information Content

The information content explored in the above adaptive estimation can be summarized as follows:

$$\begin{aligned} \mathcal{A}_2(\theta) = \{&(X_1, \tilde{X}_1, X, \tilde{X}), \phi(X_1, X, \tilde{X}_1, \tilde{X}; \theta_0) \geq 0 > \phi(X_1, x, \tilde{X}_1, \tilde{X}; \theta) \\ &\text{or } \phi(X_1, x, \tilde{X}_1, \tilde{X}; \theta_0) < 0 \leq \phi(X_1, X, \tilde{X}_1, \tilde{X}; \theta)\}. \end{aligned}$$

Then we cannot distinguish, from the true parameter θ_0, all impostors in

$$\overline{\mathcal{A}}_2 = \{\theta : P(\mathcal{A}_2(\theta)) = 0\}.$$

In a simple example, if $\text{Supp}(X_1, X) = [a, b] \times [c, d]$, then θ_0 is identified if $|\alpha_0| < b - a + |\gamma_0|(d - c)$. Recall Theorem A.1, without imposing factor structure, the necessary and sufficient condition for achieving identification is $|\alpha_0| \leq b - a$. Therefore, the blue area in the Figure below is the additional parts of parameter values that is identified with factor structure but not otherwise.

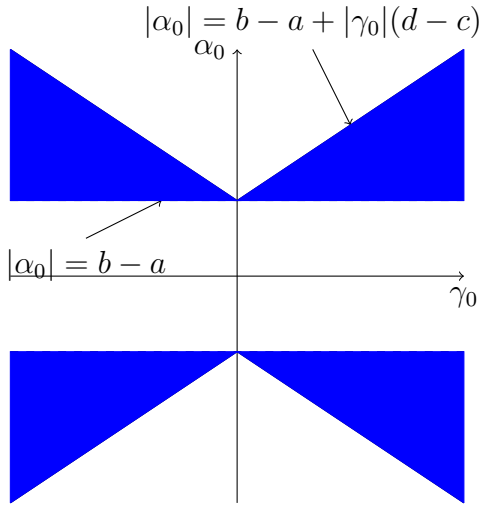

Figure 2: Identifying Power of Factor Structure

When with the factor structure, the parameter is still not identified if $|\alpha_0| > b - a + |\gamma_0|(d - c)$. In this case, if we do not impose factor structure, by Theorem A.2, the sharp identified set is $\{\alpha : \alpha > b - a\}$ while with the factor structure, the identified set (not necessarily sharp) is $|\alpha| > b - a + |\gamma|(d - c)$. This implies, when identification fails in both cases, the blue area is also the extra "bite" on the identified set given by the factor structure.

B Asymptotic Theory For Two Step Estimator

In this section we establish the asymptotic theory for the two step estimator under the conditions when the parameters are regularly identified. Many of the basic arguments follow those used in Chen and Khan (2008) and Chen, Khan, and Tang (2013). Recall what the

key identification condition that motivated the weighted least squares estimator: For pairs of observations $(x_1^{(1)}, x^{(1)})$ and $(\tilde{x}_1^{(1)}, \tilde{x}^{(1)})$ in $\text{Supp}(X_1, X)$,

$$x_1^{(1)} + \alpha_0 - \gamma_0 x^{(1)} = \tilde{x}_1^{(1)} - \gamma_0 \tilde{x}^{(1)}.$$

if and only if

$$\partial_2 P^{11}(x_1^{(1)}, x^{(1)})/f_V(x^{(1)}) + \partial_2 P^{10}(\tilde{x}_1^{(1)}, \tilde{x}^{(1)})/f_V(\tilde{x}^{(1)}) = 0.$$

Note that even though the random variable V is unobserved, the the density function $f_V(\cdot)$ above can be recovered from the data from the partial derivative of the choice probability in the selection equation with respect to the regressor in the selection equation. Thus the above equation involves the sum of two ratios of derivatives of choice probabilities.

Recall we denoted the parameter of interest by $\theta_0 \equiv (\alpha_0, \gamma_0)$.

Our estimator is based on pair of observations from the data set.We will denote the random variables of interest with capital letters, for example X_i, X_{1i}, and realizations of them with lower letters, for example x_i, x_{1i}. To denote disc tint random variables in the sample when form pairs, we will use the subscripts i, j.

Note from above, we can express the equation where the pairs receive positive weights (those whose derivatives of choice probabilities summed up to 0) as

$$X_{1i} - X_{1j} = \alpha_0 + \gamma_0(X_i - X_j) \tag{B.1}$$

So this motivates regressing the scalar random variable $X_{1i} - X_{1j}$ on the two by one vector $\tilde{X}_i \equiv (1, X_i)$. We can now see that if sufficient such pairs of observations, where the sum of the ratio of derivative of probabilities could be found to equal 0, θ_0 could be recovered as the unique solution to the system of equations corresponding to the pairs, as long as the matrix involving the terms $\tilde{X}_i - \tilde{X}_j$ satisfied a full rank condition. Of course such an approach is infeasible for two reasons. The first reason is that the probability functions, their derivatives, and hence the ratio of derivatives are unknown. The second reason is that even if these functions were known, if the probability functions are not discrete valued, such "matches" will occur with probability zero.

The first problem can be remedied by replacing the true probability function values with their nonparametric estimates. In the theory here we used a kernel estimator with kernel function $K(\cdot)$ and bandwidth H_n, whose properties are discussed below.. The second problem

can be dealt with through the use of "kernel weights" as has been frequently employed in the semiparametric literature.

Specifically, assuming that the ratio of derivatives of conditional probability functions were known, we use the following weighting function for pairs of observations:

$$\omega_{ij} = \frac{1}{h_n} k\left(\frac{P_{0i}^{k,l,r} + P_{0j}^{k,l,r}}{h_n}\right) \tag{B.2}$$

where here $P_{0i}^{k,l,r}$, $k = 0, 1, l = 0, 1$ denotes the ratio of derivatives of choice probabilities for the i^{th} observation. So, for example, $P_i^{1,1,r} = P^{11}(X_{1i}, X_i)/f_V(X_i)$. h_n is another bandwidth sequence, which converges to zero as the sample sizes increases, ensuring that in the limit, only pairs of observations with probability functions arbitrarily close to each other receive positive weight. $k(\cdot)$ is the kernel function, which is symmetric around 0, and assumed to have compact support, integrate to 1, and satisfy certain smoothness conditions discussed later on.

With the weighting matrix defined, a natural estimate of it, $\hat{\omega}_{ij}$ follows from replacing the true probability function values with their nonparametric, e.g. kernel, estimates. This suggests a weighted least squares estimator of θ_0, regressing $X_{1i} - X_{1j}$ on $\tilde{X}_j - \tilde{X}_i$, with weights $\hat{\omega}_{ij}$.

Specifically, we propose the following two stage procedure. The first stage is the kernel estimator of the ratio of derivatives of probability functions,[3] and the second stage estimator is defined as:

$$\hat{\theta} = (\sum_{i \neq j} \tau_i \tau_j \hat{\omega}_{ij} \Delta\tilde{X}_{ij} \Delta\tilde{X}_{ij}')^{-1} (\sum_{i \neq j} -\tau_i \tau_j \hat{\omega}_{ij} \Delta\tilde{X}_{ij} \Delta X_{1ij}) \tag{B.3}$$

where $\Delta X_{1ij} \equiv X_{1i} - X_{1j}$, $\Delta\tilde{X}_{ij} \equiv \tilde{X}_i - \tilde{X}_j$ and $\tau_i \equiv \tau(X_{1i}, X_i)$ is a trimming function.

We will sketch the asymptotic properties of this estimator. Here we use similar arguments to this used in Chen and Khan (2008) and keep our notation as close as possible to that used in that paper. To simplify characterizing the asymptotic properties of this estimator and the regularity conditions we impose, we first define the following functions of $P_{0i}^{k,l,r}$:

1. $f_{(P_0^{k,l,r}} = f_{P_0^{k,l,r}}(P_{0i}^{k,l,r})$, where $f_{P_0^{k,l,r}}(\cdot)$ denotes the density function of $P_{0i}^{k,l,r}$.

[3] As specified in the regularity conditions, the conditions on the bandwidth sequence are more strict than needed for the previous estimator. Specifically, they will depend on the second stage bandwidth sequence used.

2. $\mu_{\tau i} = E\left[\tau_i | P_{0i}^{k,l,r}\right]$

3. $\mu_{\tau x i} = E\left[\tau_i \tilde{X}_i | P_{0i}^{k,l,r}\right]$

4. $\mu_{\tau x x i} = E\left[\tau_i \tilde{X}_i \tilde{X}_i' | P_{0i}^{k,l,r}\right]$

Our derivation of the asymptotic properties of this estimator are based on the following assumptions:

Assumption I (Identification) The matrix:

$$M_1 = 2E\left[f_{(P_0^{k,l,r}{}_i)}(\mu_{\tau i}\mu_{\tau x x i} - \mu_{\tau x i}\mu_{\tau x i}')\right]$$

has full rank.

Assumption K (Second stage kernel function) The kernel function $k(\cdot)$ used in the second stage (to match the sum of ratios of derivatives to 0) is assumed to have the following properties:

K.1 $k(\cdot)$ is twice continuously differentiable, has compact support and integrates to 1.

K.2 $k(\cdot)$ is symmetric about 0.

K.3 $k(\cdot)$ is an eighth order kernel:

$$\begin{aligned} \int u^l k(u)du &= 0 \text{ for } l = 1,2,3,4,5,6,7 \\ \int u^8 k(u)du &\neq 0 \end{aligned}$$

Assumption H (Second stage bandwidth sequence) The bandwidth sequence h_n used in the second stage is of the form:

$$h_n = cn^{-\delta}$$

where c is some constant and $\delta \in (\frac{1}{16}, \frac{1}{12})$.

Assumption S (Order of Smoothness of Density and Conditional Expectation Functions)

S.1 The functions $P_{0i}^{k,l,r}$ are eighth order continuously differentiable with derivatives that are bounded on the support of τ_i.

S.2 The functions $f_{P_0^{k,l,r}}(\cdot)$ and $E[\tilde{x}_i|P_0^{k,l,r} = \cdot]$ have order of differentiability of 8, with eight order partial derivatives that are bounded on the support of τ_i.

The final set of assumptions involve restrictions for the first stage kernel estimator of the ratio of derivatives. This involves smoothness conditions on the propensity scores $P_{0i}^{k,l,r}$, smoothness and moment conditions on the kernel function, and rate conditions on the first stage bandwidth sequence.

Assumption PS (Order of smoothness of propensity score and regressor density functions) The functions $P_0^{k,l,r}(\cdot)$ and $f_{X_1,X}(\cdot\cdot)$ are continuously differentiable of order p, where $p > \frac{5}{2}k$.

Assumption FK (First stage kernel function conditions) $K(\cdot)$, used to estimate the choice probabilities and their derivatives is an even function, integrating to 1 and is of order $\tilde{p}$ satisfying $\tilde{p} > \frac{5}{2}\tilde{k}$. with $\tilde{k}$ denoting the dimension of X_1, X.

Assumption FH (Rate condition on first stage bandwidth sequence) The first stage bandwidth sequence H_n is of the form:

$$H_n = c_2 n^{-\gamma/k}$$

where c_2 is some constant and γ satisfies:

$$\gamma \in \left(\frac{\tilde{k}}{\tilde{p}} \left(\frac{1}{3} + \delta \right), \frac{1}{3} - 2\delta \right)$$

where δ is regulated by Assumption **H**.

Theorem B.1 *Let $\tilde{f}_i$ denote the denote the density function of the regressors used in the first stage choice probability estimation, and let $\tilde{f}_i'$ denote its derivative. Let $f(\cdot)$ denote the p.d.f. of ε_i and define the following functions of $P_{0i}^{k,l,r}$:*

$$\mathcal{G}_{k,l,i} = E\left[\tilde{X}_i (X_{1i} - \tilde{X}_i' \theta_0) | P_{0i}^{k,l,r} \right]$$

and

$$\psi_{1i} = 2\tau_i f_{P_0^{k,l,r}i} \sum_{k,l=0,1} (y_i^{k,l} \tilde{f}_i'/\tilde{f}_i - \partial_2 P_{0i}^{k,l,r}) \mathcal{G}_{k,l,i}(\mu_{\tau i}\tilde{x}_i - \mu_{\tau x i}) \tag{B.4}$$

then under Assumptions **I,K,H,S,PS,FK,FH**,

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, M_1^{-1} V_1 M_1^{-1}) \tag{B.5}$$

where

$$V_1 = E[\psi_{1i}\psi_{1i}'] \tag{B.6}$$

B.1 Nonparametric Factor Structure

Here we describe and estimator for the case where we have a nonparametric factor structure. Recall for this model we had the following relationship between unobservable variables:

$$U = g_0(V) + \bar{\Pi} \tag{B.7}$$

where we assumed that $\bar{\Pi} \perp V$.

Our goal in this more general setup is to identify and estimate both α_0 and g_0. Our identification is based on the condition that

$$x_1^{(1)} + \alpha_0 - g_0(x^{(1)}) = \tilde{x}_1^{(1)} - g_0(\tilde{x}^{(1)}).$$

if and only if

$$\partial_2 P^{11}(x_1^{(1)}, x^{(1)})/f_V(x^{(1)}) + \partial_2 P^{10}(\tilde{x}_1^{(1)}, \tilde{x}^{(1)})/f_V(\tilde{x}^{(1)}) = 0.$$

Using the same i, j pair notation as before, this gives gives us, in the nonparametric case,

$$X_{1i} - X_{1j} = \alpha_0 + (g_0(X_i) - g_0(X_j)) \tag{B.8}$$

Note the above equation has a "semi parametric form", loosely related to the model considered in, for example, Robinson (1988). However, we point out crucial differences between what we have above and the standard semi linear model. Here we are trying to identify the intercept α_0 which is usually not identified in the semi linear model as it cannot be separately identified from the nonparametric function. However, note above on the right hand side, we do not just have a nonparametric function of X_i, X_j, but the difference of two *identical* and *additively separable* functions $g_0(\cdot)$. In fact it is this differencing of these functions which enables us to separately identify α_0. which is not usually identified in semi linear models. Furthermore, as will now see when turning to our estimator of α_0, the structure of the nonparametric component, specifically additive separability of two identical functions of X_i, X_j respectively, can easily be incorporated into our approximation of each of them. From a theoretical perspective separable functions have the advantage effectively being a one dimensional problem, as there are no interaction terms to have to deal with. It is well known that nonparametric estimation of separable functions do not suffer from the "curse of dimensionality". See, for example Newey (1994).

B.1.1 Estimation of the Semilinear Model

The semi linear model, is usually expressed as

$$y_i = x_i'\beta_0 + g(z_i) + \varepsilon_i$$

where y_i denotes the observed dependent variable, x_i, z_i are observed regressors, $g(\cdot)$ is an unknown nuisance function, ε_i is an unobserved disturbance term, and β_0 is the unknown regression coefficient vector which is the parameter of interest. There is a very extensive literature in both econometrics and statistics on estimation and inference methods for this model- see for example Powell (1994) for some references.

One popular way to estimate this model is to use an expansion of basis functions, for example polynomials or splines to approximate $g(\cdot)$, and from a random sample of n observations of (y_i, x_i, z_i) regress y_i on $x_i, b(z_i)$ where $b(z_i)$ denotes the set of basis functions used to approximate $g(\cdot)$.

As an illustrative example, assuming z_i were scalar, if one were to use polynomials as basis functions, one would estimate the approximate model,

$$y_i = x_i'\beta_0 + \gamma_1 z_i + +\gamma_2 z_i^2 + \gamma_3 z_i^3 + ...\gamma_{k_n} z_i^{k_n} + u_{in}$$

where k_n is a positive integer smaller than the sample size n, and $\gamma_1, \gamma_2, ...\gamma_{k_n}$ are additional unknown parameters. This has been done by regressing y_i on $x_i, z_i, z_i^2, ...z_i^{k_n}$, and our estimated coefficient of x_i would be the estimator of β_0.

The validity of this approach has been shown in, for example, Donald and Newey (1994)

Now for our problem at hand, incorporating a nonparametric factor structure, we propose a kerne weighted least squares estimator. The weights are as they were before, assigning great weights to pairs of observations where the sum of derivatives of ratios of choice probabilities are closer to 0.

The dependent variable is identical to as before, the set of n choose 2 pairs $X_{1i} - X_{1j}$, The regressors now reflect the series approximation of $g_0(X_i) - g_0(X_j)$:

$$g_0(X_i) - g_0(X_j) \approx \gamma_1(X_i - X_j) + \gamma_2(X_i - X_j)^2 + \gamma_3(X_i - X_j)^3 + ...\gamma_{k_n}(X_i - X_j)^{k_n}$$

So now our estimator would be to regress $X_{1i} - X_{1j}$ on $1, (X_i - X_j), (X_i - X_j)^2, ...(X_i - X_j)^{k_n}$, using the same weights $\hat{\omega}_{ij}$ so the estimator of α_0 would be the coefficient on 1. Specifying the asymptotic properties of tis estimator would require additional regularity conditions, notable the rate at which the sequence of integers k_n increases with the sample size n.

We again only outline these regularity conditions here, and only to establish consistency:

As before we first define the following functions of $P_{0i}^{k,l,r}$:

1. $f_{(P_0^{k,l,r}} = f_{P_0^{k,l,r}}(P_{0i}^{k,l,r})$, where $f_{P_0^{k,l,r}}(\cdot)$ denotes the density function of $P_{0i}^{k,l,r}$.
2. $\mu_{\tau i} = E\left[\tau_i | P_{0i}^{k,l,r}\right]$
3. $\mu_{\tau x i} = E\left[\tau_i g_0(\tilde{X}_i) | P_{0i}^{k,l,r}\right]$
4. $\mu_{\tau x x i} = E\left[\tau_i g_0(\tilde{X}_i) g_0(\tilde{X}_i)' | P_{0i}^{k,l,r}\right]$

Our derivation of the asymptotic properties of this estimator are based on the following assumptions:

Assumption I2 (Identification) The matrix:

$$M_1 = 2E\left[f_{(P_0^{k,l,r}i)}(\mu_{\tau i}\mu_{\tau xxi} - \mu_{\tau xi}\mu'_{\tau xi})\right]$$

has full rank.

Assumption K2 (Second stage kernel function) The kernel function $k(\cdot)$ used in the second stage (to match the sum of ratios of derivatives to 0) is assumed to have the following properties:

K2.1 $k(\cdot)$ is twice continuously differentiable, has compact support and integrates to 1.

K2.2 $k(\cdot)$ is symmetric about 0.

K2.3 $k(\cdot)$ is a second order order kernel:

$$\begin{aligned}\int u^l k(u)du &= 0 \text{ for } l = 1\\ \int u^2 k(u)du &\neq 0\end{aligned}$$

Assumption H2 (Second stage bandwidth sequence) The bandwidth sequence h_n used in the second stage satisfies $h_n \to 0$ and $nH_nh_n^2 \to \infty$. where H_n denotes the first stage bandwidth sequence.

Assumption S2 (Order of Smoothness of Density and Conditional Expectation Functions)

S2.1 The functions $P_{0i}^{k,l,r}$ are eighth order continuously differentiable with derivatives that are bounded on the support of τ_i.

S2.2 The functions $f_{P_0^{k,l,r}}(\cdot)$ and $E[\tilde{x}_i|P_0^{k,l,r} = \cdot]$ have order of differentiability of 8, with eight order partial derivatives that are bounded on the support of τ_i.

The final set of assumptions involve restrictions for the first stage kernel estimator of the ratio of derivatives. This involves smoothness conditions on the propensity scores $P_{0i}^{k,l,r}$,

smoothness and moment conditions on the kernel function, and rate conditions on the first stage bandwidth sequence.

Assumption PS2 (Order of smoothness of propensity score and regressor density functions) The functions $P_0^{k,l,r}(\cdot)$ and $f_{X_1,X}(\cdot\cdot)$ are continuously differentiable of order p, where $p > \frac{5}{2}k$.

Assumption FK2 (First stage kernel function conditions) $K(\cdot)$, used to estimate the choice probabilities and their derivatives is an even function, integrating to 1 and is of order $\tilde{p}$ satisfying $\tilde{p} > \frac{5}{2}\tilde{k}$. with $\tilde{k}$ denoting the dimension of X_1, X.

Assumption FH2 (Rate condition on first stage bandwidth sequence) The first stage bandwidth sequence H_n is of the form: satisfies $H_n \to 0$ and $nH_n^2 \to \infty$.

Assumption BFC (Basis function conditions) The basis function approximation of the unknown factor structure function satisfies the following conditions:

BFC.1 The number of basis functions, k_n, satisfies $k_n \to \infty$ and $k_n/n \to 0$.

BFC.2 For every k_n, the smallest eigenvalue of the matrix

$$E[P_{k_n} P_{k_n}']$$

is bounded away from 0 uniformly in k_n, where

$$P_{k_n} \equiv (1, (X_i - X_j), (X_i - X_j)^2, ...(X_i - X_j)^{k_n})'$$